Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/DK04/000914

International filing date: 23 December 2004 (23.12.2004)

Document type: Certified copy of priority document

Document details: Country/Office: DK
Number: PA 2004 00586
Filing date: 07 April 2004 (07.04.2004)

Date of receipt at the International Bureau: 11 February 2005 (11.02.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

PCT/DK 2004/000914

Kongeriget Danmark

Patent application No.:	PA 2004 00586
Date of filing:	07 April 2004
Applicant: (Name and address)	Torben F. Ørntoft Helgesvej 19 DK-8230 Aabyhøj Denmark

Title: Classification of Colon Cancer

IPC: -

This is to certify that the attached documents are exact copies of the above mentioned patent application as originally filed.

Patent- og Varemærkestyrelsen
Økonomi- og Erhvervsministeriet

03 February 2005

Susanne Morsing

Susanne Morsing

Classification of Colon Cancer

Background

Colon cancers microsatellite instability status is a better marker for response to adjuvant chemotherapy with fluorouracil than tumour stage II and III. The majority of hereditary colorectal cancer cases are microsatellite instable. We investigated the possibility of classifying colon tumors based on gene expression in crude biopsies and correlated these to crude survival and investigated if the gene expression profile can also identify hereditary cases from sporadic cases.

Methods

Gene transcripts from tumour specimens were quantified using microarray technology. The tumors were clustered using unsupervised and supervised classification algorithms. Sets of genes were defined for classification of microsatellite instability status and sporadic verses hereditary microsatellite instable tumors. Real-time PCR was used to validate microarray data and to investigate platform dependency in a new independent set of 47 colorectal tumors.

Results

Unsupervised hierarchical clustering revealed that tumors were essentially separated according to microsatellite instability status. Supervised classification of the 97 tumor samples using a maximum likelihood classifier with a crossvalidation loop resulted in tree misclassification as compared to microsatellite analysis using from 106 genes and down to only seven genes. The stability of classification of colon tumors in relation to microsatellite status was tested by permutation analysis. The sensitivity for diagnosis of microsatellite stable tumors exceeded 99% with a specificity exceeding 96%. The positive and negative predictive values exceeded 95% and 98%, respectively. The classifier was demonstrated not to be platform dependent as it could successfully be reproduced by real-time PCR. This was further verified as the classifier also correctly classified 95.7% of a new independent set of 47 colorectal tumors using real-time PCR.

Based on microarray data we identified ten genes that were highly correlated with hereditary disease. Using down to two of these genes 36 of 37 microsatellite instable tumors could be correctly separated into sporadic and hereditary MSI-H colorectal tumors.

Crude survival according to microsatellite status as determined by the classifier, revealed that stage II colon receiving no adjuvant chemotherapy, that patient displaying microsatellite instability had significantly longer overall survival than patient exhibiting microsatellite stable tumors (P=0.0014). By contrast, the patient with Dukes' C tumors displaying microsatellite instability did not have a significant increase in overall survival as compared to patient exhibiting microsatellite stable tumors (P=0.55).

Conclusion

Colon cancer can be stratified into two molecular distinct groups by quantification of the transcripts of 106 genes or even down to seven genes. The two groups are highly correlated with microsatellite stable (MSS) and microsatellite instable (MSI) tumors. The 7-gene classifier clearly proved to be a strong predictor of survival in Dukes B and it can be used to select patients who need adjuvant chemotherapy, namely those classified as MSS. We demonstrate that this classification is also valid when performed by real-time PCR analysis allowing a fast diagnosis in a clinical setting. Finally, sporadic from hereditary cases in tumors exhibiting microsatellite instability can be identified based on gene expression monitoring.

Classification of Colon Cancer

Colon is the fourth most frequently diagnosed malignancy and the second most common cause of cancer death in the western world. The standard treatment of colon cancer is advised according to tumor stage. Patient with Dukes' C colon cancer receives a flurouracil-based adjuvant systemic chemotherapy in addition to surgical resection of the tumor, whereas the treatment for Dukes' B patients is based alone on surgical resection.

There is accumulating evidence that these cancers belong to two distinct molecular types according to genetic alterations. The mutator phenotype featuring tumors with microsatellite instability (MSI) and the suppressor pathway displying chromosomal instability and microsatellite stable (MSS). MSI has been defined as a change of any length due to either insertions or deletions of repeating units in a microsatellite within a tumor compared to normal tissue and is caused by an underlying defect in the mismatch repair (MMR) system. (Boland et al, CR 1998, 58:5248). The MSI pathway may either be sporadic or hereditary (HNPCC) and whereas the disruption of the MMR system in sporadic MSI tumors is most often caused by somatic methylation of the MLH1 gene promoter more that 90% of HNPCC cancers are caused by germline mutations in MLH1 or MSH2.

The MSS pathway to cancer begins with the inactivation of tumor suppressor genes, such as APC/β-catenin genes, followed by activation of oncogenes and inactivation of additional tumor suppressor genes, commonly with a high frequency of allelic losses and cytogenetic abnormalities and abnormal DNA tumor content. Many studies have defined the pathoclinical trait of MSI and MSS tumors and found that MSI positive cancers are most frequently found in the right side of the colon, they tend to be of less differentiated, they tend to be larger in size, are often mucinous and often exhibit extensive infiltration by lymphocytes.

Crude survival data suggest that patients with HNPCC have a better prognosis than those with sporadic disease [48,49.50] and studies have also shown that MSI is an independent indicator of broad spectrum of tumors in relation to location, heredity, microsatellite instability status, and origin of the patient. All tumors were collected in the period from 1994 to 2002. 68 tumor samples were collected at nine different clinics in Finland and 33 samples were collected at four different clinics in Denmark, 36 were Dukes' B, 67 Dukes' C, 41 were sporadic microsatellite highly instable (MSI-H) of which were 17 HNPCC, and 59 were sporadic microsatellite stable (MSS) (table 1). None of the patients received pre-operative radiation or chemotherapy.

Microsatellite-instability analysis. From all tumor samples available as paraffin blocks, ten sections were cut at 10μm and stained with haematoxylin. The first and last section was cut at 4 μm and stained with haematoxylin. These two sections were used for the identification of tumor and normal cells from each sample. Regions enriched in tumor cells (more than 90%) were microdissected from these sections and DNA was extracted using a Puregene DNA extraction kit (Gentra Systems, Minneapolis, MN). DNA from blood samples was used as control when available, otherwise normal tissue was microdissected from the tissue sections. The samples were analysed for microsatellite instability according to the NCI guidelines (Boland et al). Samples positive for markers BAT25 and BAT26 were scored as MSI-H. Samples positive for only one of these markers were tested for further markers and scored as MSI-L if none of these tested positive. Since MSI-L has similar clinical features as MSS these samples were considered as MSS in this study. In addition to microsatellite analysis all tumors from which paraffin blocks were available were tested for the presence of MLH1 and MSH2 protein by immunohistochemistry. None of the samples scored MSS were negative for either protein whereas six of the MSI scored samples were positive for both (Table 1).

Classification of Colon Cancer

RNA purification Colon specimens were obtained fresh from surgery and were immediately snap frozen in liquid nitrogen either as was, in OCD-compound or in a SDS/guadinium thiocyanate solution. Total RNA was isolated using RNAzol (WAK-Chemie Medical) or spin column technology (Sigma) following the manufactures' instructions.

Gene expression analysis These procedures were performed at described in detail elsewhere (Dyrskødt et al). Briefly, ten µg of total RNA was used as starting material for the target preparation as described. First and second strand cDNA synthesis was performed using the SuperScript II System (Invitrogen) according to the manufacturers' instructions except using an oligo-dT primer containing a T7 RNA polymerase promoter site. Labelled aRNA was prepared using the BioArray High Yield RNA Transcript Labelling Kit (Enzo) using Biotin labelled CTP and UTP (Enzo) in the reaction together with unlabeled NTP's. Unincorporated nucleotides were removed using RNeasy columns (Qiagen). Fifteen µg of cRNA was fragmented, loading onto the Affymetrix HG_U133A probe array cartridge and hybridized for 16h. The arrays were washed and stained in the Affymetrix Fluidics Station and scanned using a confocal laser-scanning microscope (Hewlett Packard GeneArray Scanner G2500A). The readings from the quantitative scanning were analyzed by the Affymetrix Gene Expression Analysis Software (MAS 5.0) and normalized using RMA (robust multi array normalisation, Irizarry et al. 2002) in the statistical application R. Redundant probesets (as defined form Unigene build 168) with high correlation (>0.5) over all samples were removed, which reduced the dataset to approximately 14.400 probesets. This dataset was used a source for all further calculations in this manuscript.

Classification of Colon Cancer

Unsupervised agglomerative hierarchical clustering

For hierarchical cluster analysis 1239 genes with a variation across all samples greater than 0.5 were median-centred to a magnitude of 1. Samples and genes were then clustered using average linkage clustering with a modified Person correlation as similarity metric (Eisen et al., PNAS 95: 14863-14868, 1998). The cluster dendrogram was visualized with TreeView (Eisen).

Group testing

We make a statistical test where the p-value is evaluated through permutations. For each group and gene we calculate the average and the sum of squared deviations from the average. We then sum these over the genes and the groups:

$$\boxed{S_1 = \sum_{\text{groups}} \sum_{\text{genes}} (X_{ij} - \bar{X}_{gr(i)j})^2}$$

This expression is calculated for joining DK with SF and MSI with MSS such that we end up with two groups. The sum of squared deviations is denoted S_2. As a test statistic we use S_1/S_2. A small value indicates that there is a real reduction in the deviations when going from 2 to 4 groups and thus the groups have a real significance. To judge if a value is significantly small we use permutations. For each of the four groups left when joining DK and SF we randomly allocate the members to a pseudo DK and pseudo SF in such a way that the number of members in each group are as in the original data

To get an understanding of this separation we performed a test to see if this is caused by few genes or if many genes are involved. For this test we calculated $S_1 = \sum_{\text{genes}} S_1(\text{gene})$ and similarly with $S_2 = \sum_{\text{genes}} S_2(\text{gene})$. For each gene j we used the test statistic $S_1(j)/S_2(j)$ (Table 3).

Classification of Colon Cancer

Multidimentional Scaling

We carried out multidimentional scaling on median-centered and normalized data using CMD—scale in the statistical application R and visualized in a two-dimentional plot.

Microsatellite status classifier

The readings from the quantitative scanning were analyzed by the Affymetrix Gene Expression Analysis Software (MAS 5.0) and normalized using RMA (robust multi array normalisation. Irizarry et al. 2002) in the statistical application R. Redundant probesets (as defined form Unigene build 168) with high correlation (>0.5) over all samples were removed, which reduced the dataset to approximately 14.400 probesets.

The microsatellite instability status classifier was based on a dataset of 4.266 genes. These genes result from the removal of genes with a variance over all tumor samples smaller than 0.2 and genes that separate Danish from Finnish samples with a t-value numerically greater than 2. We used a normal distribution with the mean dependent on the gene and the group (MSI, MSS). For each gene, we calculated the variation between the groups and the variation within the groups to select genes with a high ratio between these. To classify a sample, we calculated the sum over the genes of the squared distance from the sample value to the group mean, standardized by the variance and assigned the sample to the nearest group. The sample to be classified was excluded when calculating group means and variances.

Estimation of classifier stability

We validated the performance of the classifier by permutation. One hundred datasets consisting of 30 MSS samples and 25 MSI samples were randomly chosen by permutation for training of the classifier with the remaining samples in each case being assign to a testset. Averages over the 100 data sets of the number of errors in the cross-validation of the training set and in the test set were used as a measure of the precision of the classifier.

Real-time PCR (RT-PCR). The procedures were as described (Birkenkamp-Demtroder) except that we used short LNA (Locked Nucleic Acid) enhanced probes from a Human Probe Library (Exiqon™). In short, cDNA was synthesized from single samples some of which were previously analyzed on GeneChips. Reverse transcription was performed using Superscript II RT (Invitrogen). Real-time PCR analysis was performed on selected genes using the primers (DNA Technology) and probes (Exiqon, DK) described in figure legend X. All samples were normalized to GAPDH as described previously (Birkenkamp-Demtroder et. al. Cancer Res.. *62*: 4352-4363, 2002).

Rebuilding of Classifier based on Real-Time PCR

The 79 tumors samples that were not analysed by real-time PCR were transformed into log ratios using one of the tumor samples as reference and used for training of the classifier. Then 23 samples of which 18 were also analyzed on arrays were equally transformed into log ratios using the same tumor sample as above as reference and tested. The idea behind this translation is that we expect the normalized PCR values to be proportional to the normalized array values, and on a log scale this becomes an additive difference. The difference is gene specific and is therefore estimated for each gene separately. The variation obtained from the microarray data, and used in the classifier, can be used directly on the PCR platform.

Results

Hierarchical Clustering

The clinical specimens used in this study were collected in two different countries from 14 different clinics in the period 1994 to 2001. The samples were selected to keep a balanced representation of microsatellite instable (MSI) and microsatellite stable (MSS) tumors from both the right- and left-sided colon. The MSI class was represented both by sporadic MSI and hereditary MSI (HNPCC) tumors. Only Dukes' B and Dukes' C tumor samples were included were selected (table 1). Before any attempt to divide a diverse sample collection into distinct classes analyzed the data for systematic bias that may have been introduces during the experimental procedures. A fast and easy way to discover both true distinct classes as well as systematic biases in the data is to perform a hierarchical clustering.

The phylogenetic tree resulting from hierarchical clustering on 1239 genes (fig 1) reveals that the main separating factor is microsatellite status. On the upper trunk we find two clusters represented mainly by normal biopsies (14/21) and MSS tumors (18/25), respectively. The lower trunk is divided into a MSI cluster (30/36) and a second MSS cluster (MSS2-cluster) (34/37). A closer inspection of the two MSS clusters unveil that one is dominated by Danish samples (19/25) and one by Finnish samples (26/37 check). Also, it is worth to notice that the MSI cluster contains a vast majority of Finnish samples (32/36) and that the sporadic MSI samples are interspersed among the hereditary samples. The normal biopsies cluster tight together with a slight tendency to separation according to origin. Tree normal samples cluster within the MSI cluster indicating that resection of these samples may have been to close to the tumor lesion.

Inspection of the gene cluster dendrogram shows that the two groups of MSS tumors are mainly separated by a large cluster of genes being upregulated in the Danish samples (data not shown) indicating that a systematic difference between Danish and Finnish samples.

Significance of Observed Groups

Based on these observations, we performed a series of test to evaluate if the observed separation of tumors into MSS and MSI as well as DK and SF are significant. For these tests the tumor samples were grouped into four virtual tumor-groups labelled, i.e. Danish MSI (MSI-DK), Danish MSS (MSS-DK), Finnish MSI (MSI-SF) and Finnish MSS (MSS-SF). Based on 5082 genes with a variance above 0.2, we tested if all four groups are significant or if some of the groups can be joined. We considered the two possibilities of joining DK and SF, and of joining MSI and MSS and made a statistical test where the p-value is evaluated through permutations. In 100 permutations of each group combination our test value S1/S2 is considerably smaller than in all permutation (Table 2) demonstrating a very clear separation between DK and SF and between MSI and MSS. Such a clear distinction between groups may rely on a few highly separating genes or a general difference in the gene expression profile including many genes. For both the DK-SF and MSI-MSS the effect are caused by many genes even at very criteria, i.e. low test statistic $S_1(j)/S_2(j)$ values (Table 3).

When a property is present that influences a large proportion of the genes this may obscure separation of clinical relevant features in unsupervised clustering. To visualize the effect of such properties, we calculated distances by multidimensional scaling between samples with and without of 816 genes separating DK from SF with a t-value numerically greater than 2 (Fig 2). We see an improved separation of MSI and MSS with Danish and Finnish cases mixed. The MSI-DK samples are not completely separated as they are found both between the MSI-SF and the MSS samples. (These plots are not entirely unsupervised since the groups have been used to remove gene).

Construction of an MSI-MSS classifier

For the construction of a classifier we used the expression profiles from 97 tumors for which no ambiguity had been identified in relation to microsatellite status. The 816 genes separating DK from SF were excluded, as these would be unreliable for MS classification. We built a maximum likelihood classifier in order to select a minimum of genes giving the largest possible separation of the two groups. We tested the performance of the classifier using 1-1000 genes and found that it was stable showing 3-6 errors when using 4 – 400 genes. Of these 106 genes were especially suited for discrimination of MSS from MSI (table 4). The minimum of three errors was found even using only 7 genes (Table 5).

Classification of ambiguous samples

Application of the 7-gene classifier to the four samples showing ambiguity in the microsatellite analyses assigns all four to be microsatellite stable tumor class. Notably, all four showed expression levels of *Tumor Growth Factor β induced protein* (TFGBI). MLH1 and thymidylate synthase (TYMS) that are atypical for MSI tumors. Furthermore, these tumors were all from the left colon. Thus the misclassified tumors are clearly truly MSS or they belong to a yet undefined class of MSI tumors.

Stability of classification

To estimate the stability of the classifier based on all 97 tumor samples, we generated one hundred new classifiers based on randomly chosen datasets consisting of 30 MSS and 25 MSI samples. In each case the classifiers were tested with the remaining samples. The performance for each set was evaluated and averaged over all 100 training and test sets (Table 6). The mean error rate for MSS tumors was 0.52% and 1.38% for MSI tumors. The seven genes defined above were found to be those genes that were most frequently used in the crossvalidation loop. More than 50% of the errors were related to three tumors of which two were wrongly classified in all permutation and one in 94%. The remaining errors were mainly caused by four tumors with error rates of 40-47% showing that the former three samples are truly assigned contradictory to result from the microsatellite analysis and that four samples could not be assigned with confidence too any of the classes.

Survival classifier

Using the same classification methods described above, we build classifiers for survival based on either all samples or the above defined groups of MSI-H and MSS. As seen in figure 3, a distinction of patient with good prognosis (>5 year survival) from patient with bad prognosis (<5 years survival) can be achieved with higher precision and using only a fraction of the genes by first separating into MSI-H and MSS groups.

Construction of a classifier for sporadic versus hereditary microsatellite instable tumors

In order to identify a gene set for identification of hereditary microsatellite instable tumors we applied 19 sporadic microsatellite instable samples and 18 microsatellite instable samples to supervised classification as described above. We found ten genes we high scored for separation of sporadic MSI-H from hereditary MSI-H tumours (Table 8). In crossvalidation we found a minimum number of one error using two genes (Fig 4A) and were used in at least 36 of the 37 crossvalidation loops. The genes were: the mismatch repair gene MLH1 that show a general downregulation in sporadic disease and PIWIL1 that is lower expressed in hereditary cases (Fig 4B). Using these two genes only one error occurred: a sporadic microsatellite instable was classified as hereditary. Based on T-test we performed 500 permutations to test the significance of these two genes for marker genes and found both genes highly significant with p-values < 0.005.

Cross platform classification

Real time PCR was applied both to verify the array data and examine if the 7-gene classifier would also perform on this platform. We chose 23 samples of which 18 were also analyzed on arrays. The correlation between the two platforms was high (data not shown). In order to test the performance of classification using PCR data we re-build our classifier with a 79 samples array dataset including only those tumors that were not analyzed with PCR. Two samples were classified in discordance with the microsatellite instability test of which one of them was ambiguously classified by the 7-gene array classifier.

Relation between microsatellite-instability status, stage and survival

Based on the 7-gene classifier. classification of 36 patients with Dukes' B tumors receiving no adjuvant chemotherapy, 18 were classified as MSI tumors and 18 as MSS tumors. The overall survival was highly significantly related to the classification since all nine patients that died within five years of follow-up were belonged to the MSS group (P=0.0014) (Fig. 5A). Thus, the 7-gene classifier clearly proved to be a strong predictor of survival in Dukes B and it can be used to select patients who need adjuvant chemotherapy, namely those classified as MSS.

Classification of Colon Cancer

Among 65 patients with Dukes' C tumors receiving adjuvant chemotherapy, 17 were classified as MSI tumors and as 48 MSS tumors. Of these, 6 MSI and 27 MSS patients died within five years of follow-up meaning no significant difference in overall survival between these groups (P=0.55) (Fig.5B). A trend was that the MSI showed a poorer short-term survival than the MSS, contrary to Dukes B patients. This difference can be attributed to the fact that a recent large study has shown that chemotherapy only benefit the MSS tumor patients, thus improving their survival to a level comparable to that which is characteristic of MSI tumor patients.

Clinical application of the discovery

In the clinic the 106 or less genes described can be used for predicting outcome of colorectal cancer when examined at the RNA level and also on the protein level as each gene identified is the project is transcribed to RNA that is further translated into protein. The genes can also be used determine which patient should be treated with chemotherapy as only non-microsatellite instable tumors will respond to 5-FU based therapy. Building classifiers can achieve a further stratification of patient with god and bad prognosis after stratification into microsatellite instable and stable tumors. The genes used to identify hereditary disease can be used to decide which patient should enter into sequencing analysis of mismatch repair genes.

The RNA determination can be made in any form using any method that will quantify RNA. The proteins can be measured with any method quantification method that can determine the level of proteins.

References

Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D, Yeatman TJ. Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. J Natl Cancer Inst. 2002 Apr 3;94(7):513-21.

Classification of Colon Cancer

Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, Laurberg S, Sorensen FB, Hagemann R, ORntoft TF. Gene expression in colorectal cancer. Cancer Res. 2002 Aug 1;62(15):4352-63.

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 1998 Nov 15;58(22):5248-57. Review.

Chapusot C, Martin L, Bouvier AM, Bonithon-Kopp C, Ecarnot-Laubriet A, Rageot D, Ponnelle T, Laurent Puig P, Faivre J, Piard F. Microsatellite instability and intratumoural heterogeneity in 100 right-sided sporadic colon carcinomas. Br J Cancer. 2002 Aug 12;87(4):400-4.

Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF. Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet. 2003 Jan;33(1):90-6.

Frederiksen CM, Knudsen S, Laurberg S, Orntoft TF. Classification of Dukes' B and C colorectal cancers using expression arrays. J Cancer Res Clin Oncol. 2003 May;129(5):263-71.

Huang J, Qi R, Quackenbush J, Dauway E, Lazaridis E, Yeatman T. Effects of ischemia on gene expression J Surg Res. 2001 Aug;99(2):222-7.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15;31(4):e15.

Loukola A, Eklin K, Laiho P, Salovaara R, Kristo P, Jarvinen H, Mecklin JP, Launonen V, Aaltonen LA. Microsatellite marker analysis in screening for hereditary nonpolyposis colorectal cancer (HNPCC). Cancer Res. 2001 Jun 1;61(11):4545-9.

Markowitz S, Hines JD, Lutterbaugh J, Myeroff L, Mackay W, Gordon N, Rustum Y, Luna E, Kleinerman J. Mutant K-ras oncogenes in colon cancers Do not predict Patient's chemotherapy response or survival. Clin Cancer Res. 1995 Apr;1(4):441-5.

Mori Y, Selaru FM, Sato F, Yin J, Simms LA, Xu Y, Olaru A, Deacu E, Wang S, Taylor JM, Young J, Leggett B, Jass JR, Abraham JM, Shibata D, Meltzer SJ. The impact of microsatellite instability on the molecular phenotype of colorectal tumors. Cancer Res. 2003 Aug 1;63(15):4577-82.

Classification of Colon Cancer

Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, Tu D, Redston M, Gallinger S. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer.
N Engl J Med. 2003 Jul 17;349(3):247-57.

Figure Title.

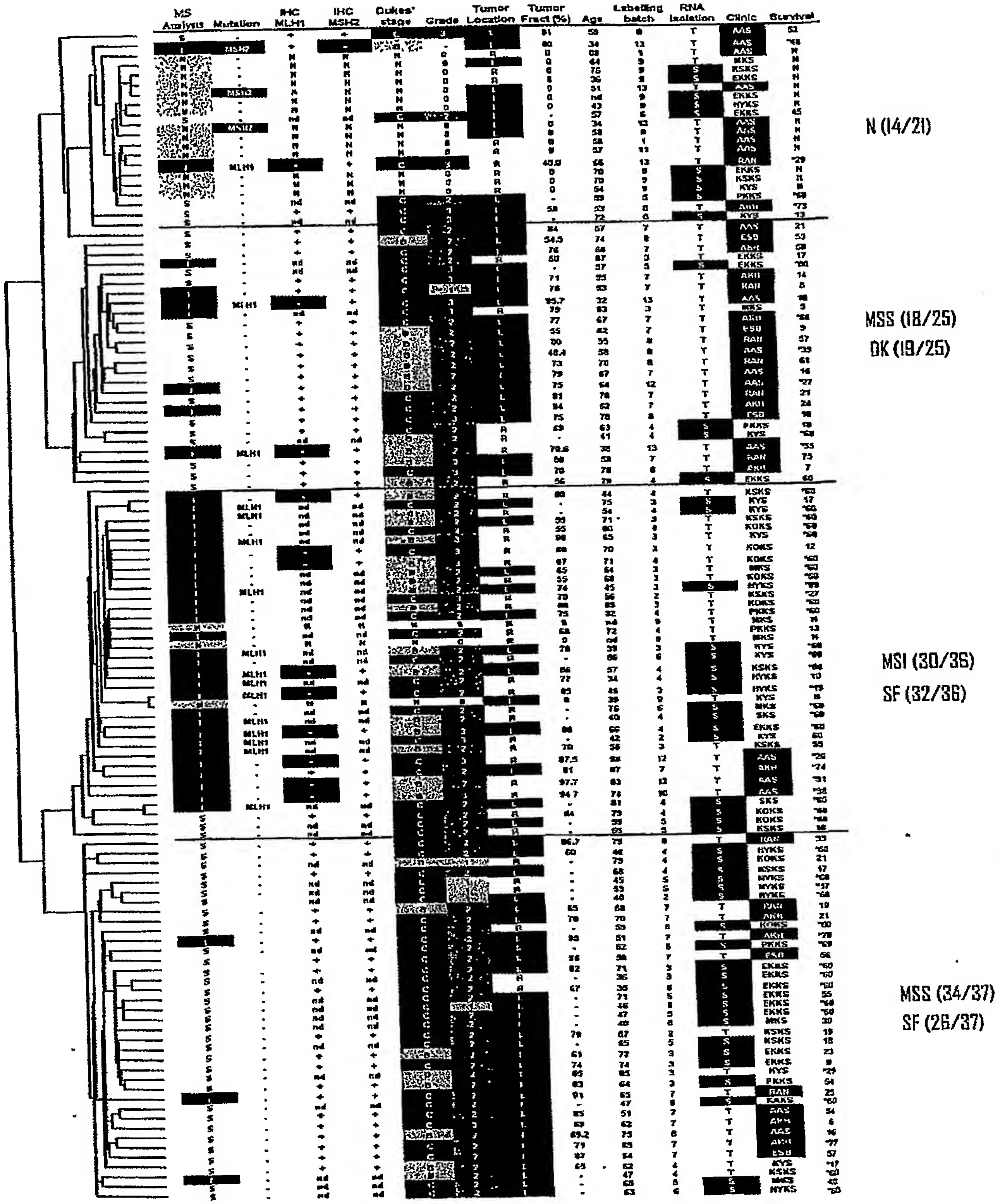


Figure 1. Clusteranalysis of Colon Specimens with Associated Clinicopathological Features.

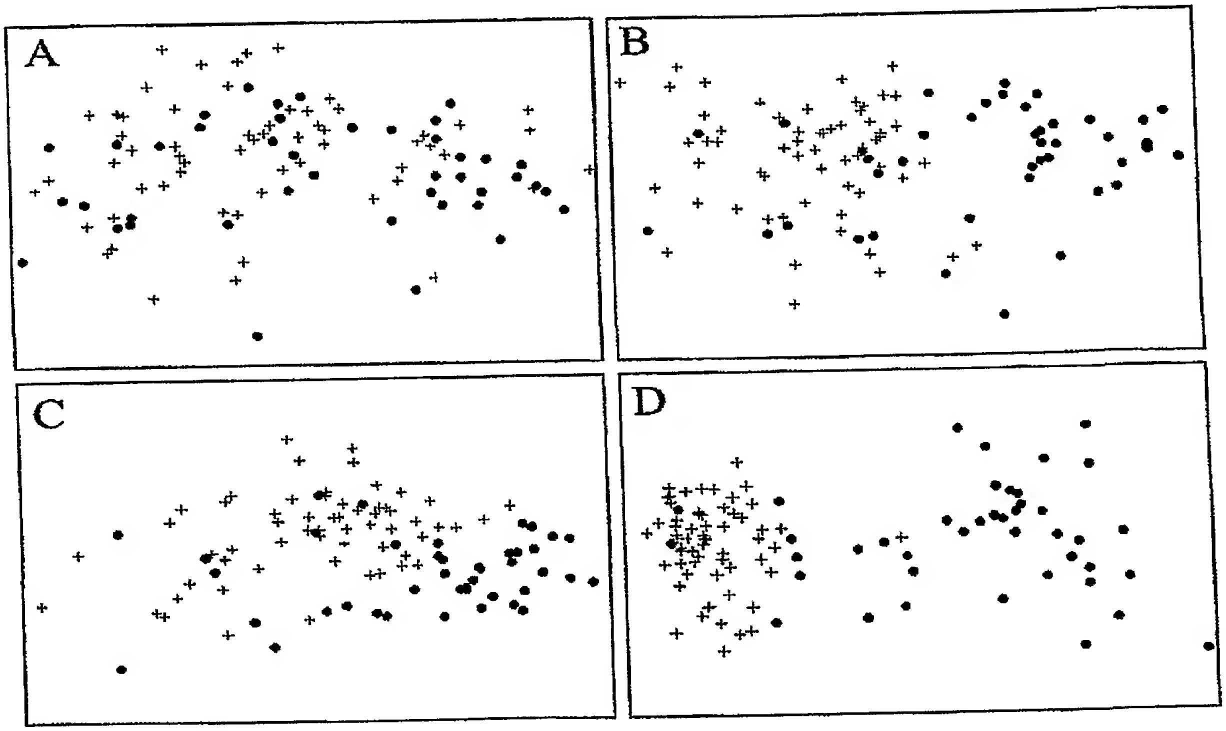
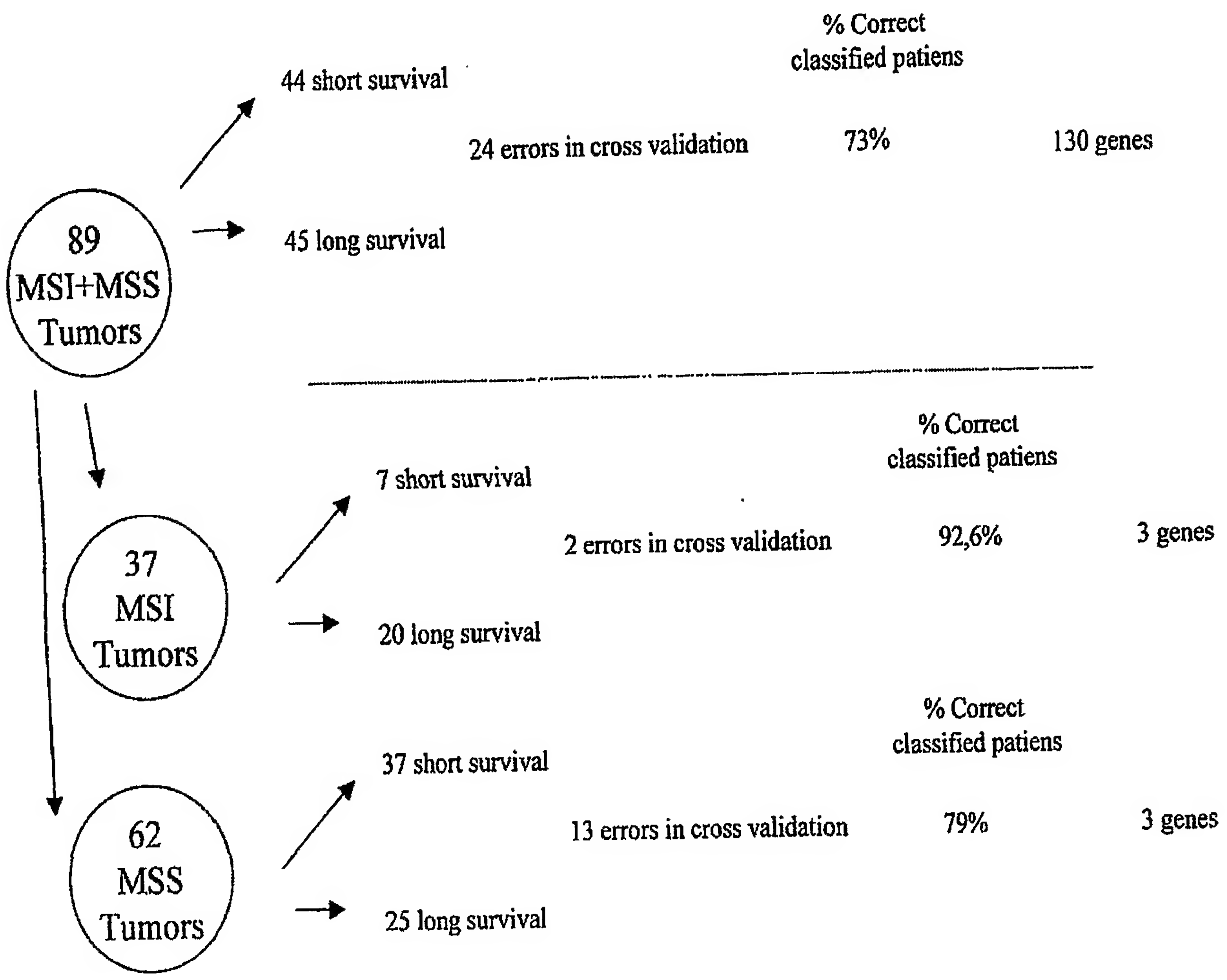


Figure 2. Multidimentional Analysis showing distances between groups of tumors.

Figure 3


89
MSI+MSS
Tumors
44 short survival
45 long survival
24 errors in cross validation
% Correct
classified patiens
73%
130 genes
37
MSI
Tumors
7 short survival
20 long survival
2 errors in cross validation
% Correct
classified patiens
92,6%
3 genes
62
MSS
Tumors
37 short survival
25 long survival
13 errors in cross validation
% Correct
classified patiens
79%
3 genes

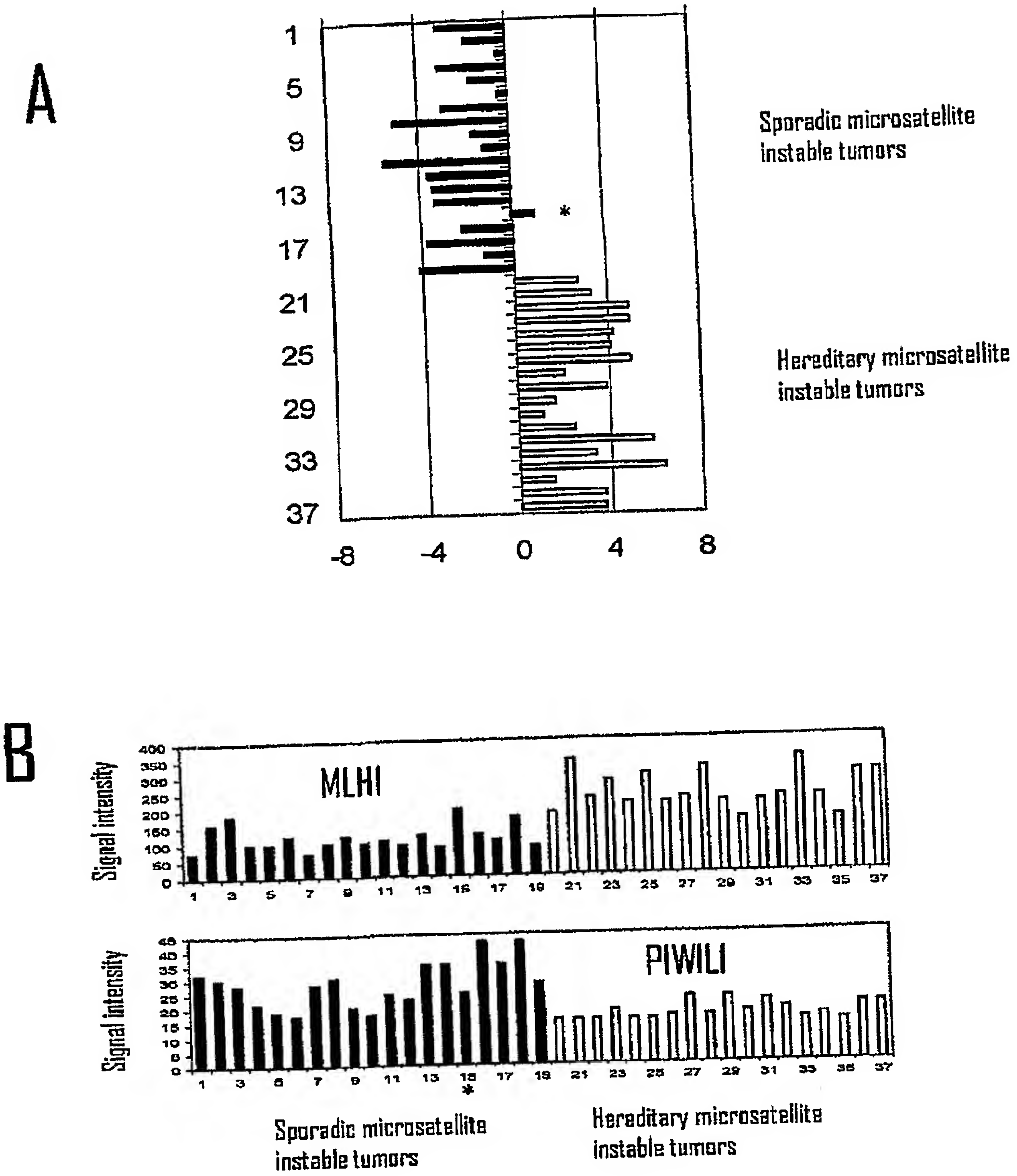
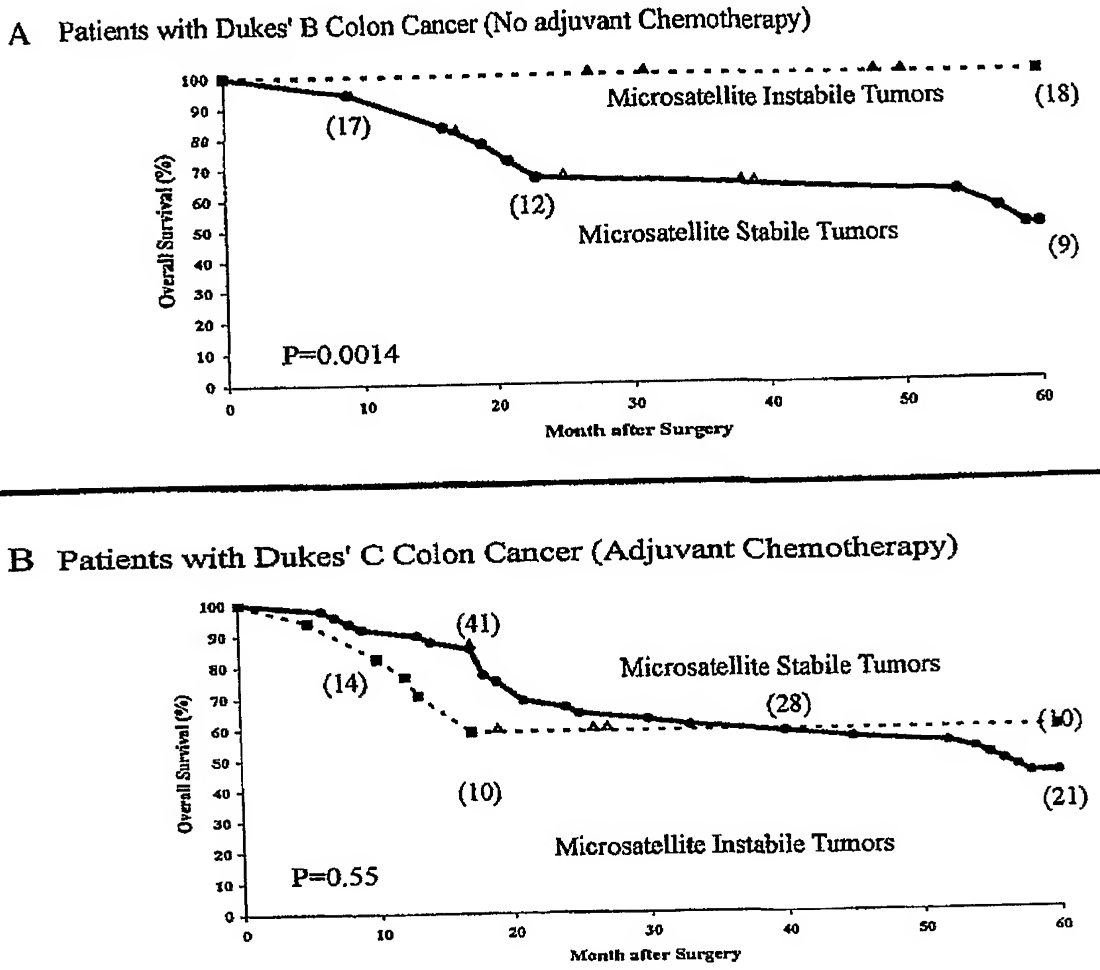

A
1
5
9
13
17
21
25
29
33
37
-8
-4
0
4
8
*
Sporadic microsatellite instable tumors
Hereditary microsatellite instable tumors
B
Signal intensity
MLHI
400
350
300
250
200
150
100
50
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
Signal intensity
PIWILI
45
40
35
30
25
20
15
10
5
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
*
Sporadic microsatellite instable tumors
Hereditary microsatellite instable tumors

Figure 4



A Patients with Dukes' B Colon Cancer (No adjuvant Chemotherapy)

B Patients with Dukes' C Colon Cancer (Adjuvant Chemotherapy)

Figure 5. Kaplan-Meier Estimates of Overall Survival among Patients with Dukes' B and Dukes' C Colon Cancer According to the Microsatellite-Instability Status of the Tumor.

Table 1. Summary of clinocopathological and microsatellite features of colon samples

Patient group		Median age	Localization in colon		Dukes' Stage			IHC negative stain	
					n (DK,SF)			N (n tested)	
n (DK,SF)		range	right (DK,SF)	left (DK,SF)	N[a]	B	C	MLH1	MSH2
All cases	119 (44,75)	62.0	45 (8,37)	74 (36,38)	17 (6,11)	35 (14,22)	66 (20,46)	12 (56)	1 (56)
MSI-H[b]	24 (8,16)	67.0	15 (3,12)	9 (6,4)	-	10 (3,7)	14 (5,9)	6 (11)	0 (11)
HNPCC[c]	17(4,13)	45.0	9 (2,7)	8 (2,6)	-	10 (2,8)	7 (2,5)	6 (8)	1 (8)
MSS	60 (25,35)	63.0	11 (0,11)	49 (25,24)	-	16 (9,7)	44 (16,28)	0 (37)	0 (37)

[a] normal biopsy taken from the resection edge of a tumor

[b] according to microsatellite analysis

[c] all tumors MSI-H

Table 2. Permutation test of groups

Pseudo group	S1/S2 from data	Smaller values in 100 permutations	Minimum in 100 permutations
DK-SF	0.9072795	0	0.962269
I-S	0.9166195	0	0.9583325

Table 3. Permutation test of genes

		$S_1(j)/S_2(j)$			
Pseudo group		< 0.6	< 0.7	< 0.8	< 0.9
DK-SF	number of genes	36	136	522	1785
	max in 100 permutations	0	0	2	225
MSI-MSS	number of genes	17	103	399	1507
	max in 100 permutations	0	1	8	250

AFFYID	SYMBOL	LOCUSLINK	OMIM	REFSEQ	GENENAME
1405_i_at	CCL5	6352	187011	NM_002985	chemokine (C-C motif) ligand 5
200628_s_at	WARS	7453	191050	NM_004184	tryptophanyl-tRNA synthetase
200814_at	PSME1	5720	600654	NM_006263, NM_006263	proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)
201641_at	BST2	684	600534	NM_004335	bone marrow stromal cell antigen 2
201649_at	UBE2L6	9246	603890	NM_004223, NM_004223	ubiquitin-conjugating enzyme E2L 6
201674_s_at	AKAP1	8165	602449	NM_003488, NM_003488	A kinase (PRKA) anchor protein 1
201762_s_at	PSME2	5721	602161	NM_002818	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)
201884_at	CEACAM5	1048	114890	NM_004363	carcinoembryonic antigen-related cell adhesion molecule 5
201910_at	FARP1	10160	602654	NM_005766	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)
201976_s_at	MYO10	4651	601481	NM_012334	myosin X
202072_at	HNRPL	3191	603083	NM_001533	heterogeneous nuclear ribonucleoprotein L
202203_s_at	AMFR	267	603243	NM_001144, NM_001144	autocrine motility factor receptor
202262_x_at	DDAH2	23564	604744	NM_013974	dimethylarginine dimethylaminohydrolase 2
202510_s_at	TNFAIP2	7127	603300	NM_006291	tumor necrosis factor, alpha-induced protein 2
202520_s_at	MLH1	4292	120436	NM_000249	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
202589_at	TYMS	7298	188350	NM_001071	thymidylate synthetase
202637_s_at	ICAM1	3383	147840	NM_000201	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
202678_at	GTF2A2	2958	600519	NM_004492	general transcription factor IIA, 2, 12kDa
202762_at	ROCK2	9475	604002	NM_004850	Rho-associated, coiled-coil containing protein kinase 2
203008_x_at	APACD	10190		NM_005783	ATP binding protein associated with cell differentiation
203315_at	NCK2	8440	604930	NM_003581	NCK adaptor protein 2
203335_at	PHYH	5264	602026	NM_006214	phytanoyl-CoA hydroxylase (Refsum disease)
203444_s_at	MTA2	9219	603947	NM_004739	metastasis-associated gene family, member 2
203559_s_at	ABP1	26	104610	NM_001091	amiloride binding protein 1 (amine oxidase (copper-containing))
203773_x_at	BLVRA	644	109750	NM_000712	biliverdin reductase A
203896_s_at	PLCB4	5332	600810	NM_000933, NM_000933	phospholipase C, beta 4
203915_at	CXCL9	4283	601704	NM_002416	chemokine (C-X-C motif) ligand 9
204020_at	PURA	5813	600473	NM_005859	purine-rich element binding protein A
204044_at	QPRT	23475	606248	NM_014298	quinolinate phosphoribosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))
204070_at	RARRES3	5920	605092	NM_004585	retinoic acid receptor responder (tazarotene induced) 3
204103_at	CCL4	6351	182284	NM_002984	chemokine (C-C motif) ligand 4
204131_s_at	FOXO3A	2309	602681	NM_001455	forkhead box O3A
204326_x_at	MT1X	4501	156359	NM_005952	metallothionein 1X
204415_at	G1P3	2537	147572	NM_002038, NM_002038, NM_022873	interferon, alpha-inducible protein (clone IFI-6-16)
204533_at	CXCL10	3627	147310	NM_001565	chemokine (C-X-C motif) ligand 10
204745_x_at	MT1G	4495	156353	NM_005950, NM_005950	metallothionein 1G
204780_s_at	TNFRSF6	355	134637	NM_000043, NM_000043, NM_152877, NM_152876, NM_152875, NM_152872, NM_152873, NM_152871	tumor necrosis factor receptor superfamily, member 6
204858_s_at	ECGF1	1890	131222	NM_001953	endothelial cell growth factor 1 (platelet-derived)
205241_at	SCO2	9997	604272	NM_005138	SCO cytochrome oxidase deficient homolog 2 (yeast)
205242_at	CXCL13	10563	605149	NM_006419	chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)
205495_s_at	GNLY	10578	188855	NM_006433, NM_006433	granulysin
205831_at	CD2	914	186990	NM_001767	CD2 antigen (p50), sheep red blood cell receptor
206108_s_at	SFRS6	6431	601944	NM_006275	splicing factor, arginine/serine-rich 6
206286_s_at	TDGF1	6997	187395	NM_003212	teratocarcinoma-derived growth factor 1
206461_x_at	MT1H	4496	156354	NM_005951	metallothionein 1H
206754_s_at	CYP2B6	1555	123930	NM_000767	cytochrome P450, family 2, subfamily B, polypeptide 6
206907_at	TNFSF9	8744	606182	NM_003811	tumor necrosis factor (ligand) superfamily, member 9
206918_s_at	RBM12	10137	607179	NM_006047, NM_006047	RNA binding motif protein 12
206976_s_at	HSPH1	10808		NM_006644	heat shock 105kDa/110kDa protein 1
207320_x_at	STAU	6780	601716	NM_004602, NM_004602, NM_017452, NM_017453	staufen, RNA binding protein (Drosophila)
207457_s_at	LY6G6D	58530	606038	NM_021246	lymphocyte antigen 6 complex, locus G6D
207993_s_at	CHP	11261	606988	NM_007236	calcium binding protein P22
208022_s_at	CDC14B	8555	603505	NM_003671, NM_003671, NM_033331	CDC14 cell division cycle 14 homolog B (S. cerevisiae)
208156_x_at	EPPK1	83481			epiplakin 1
208581_x_at	MT1X	4501	156359	NM_005952	metallothionein 1X
208944_at	TGFBR2	7048	190182	NM_003242	transforming growth factor, beta receptor II (70/80kDa)
209048_s_at	PRKCBP1	23613		NM_012408, NM_012408, NM_183047	protein kinase C binding protein 1
209108_at	TM4SF6	7105	300191	NM_003270	transmembrane 4 superfamily member 6
209504_s_at	PLEKHB1	58473	607651	NM_021200	pleckstrin homology domain containing, family B (evectins) member 1
209546_s_at	APOL1	8542	603743	NM_003661, NM_003661, NM_145343	apolipoprotein L, 1
210029_at	INDO	3620	147435	NM_002164	indoleamine-pyrrole 2,3 dioxygenase

Table 4.

210103_s_at	FOXA2	3170	600288	NM_021784, NM_021784	forkhead box A2
210321_at	GZMH	2999	116831	NM_033423	granzyme H'(cathepsin G-like 2, protein h-CCPX)
210538_s_at	BIRC3	330	601721	NM_001165, NM_001165	baculoviral IAP repeat-containing 3
211456_x_at	AF333388				
212057_at	KIAA0182	23199		XM_050495	KIAA0182 protein
212070_at	GPR56	9289	604110	NM_005682	G protein-coupled receptor 56
212185_x_at	MT2A	4502	156360	NM_005953	metallothionein 2A
212229_s_at	FBXO21	23014		NM_015002, NM_015002	F-box only protein 21
212336_at	EPB41L1	2036	602879	NM_012156, NM_012156	erythrocyte membrane protein band 4 1-like 1
212341_at	MGC21416	286451		NM_173834	hypothetical protein MGC21416
212349_at	POFUT1	23509	607491	NM_015352, NM_015352	protein O-fucosyltransferase 1
212859_x_at	MT1E	4493	156351	NM_175617	metallothionein 1E (functional)
213201_s_at	TNNT1	7138	191041	NM_003283, NM_003283, XM_352926	troponin T1, skeletal, slow
213385_at	CHN2	1124	602857	NM_004067	chimerin (chimaerin) 2
213470_s_at	HNRPH1	3187	601035	NM_005520	heterogeneous nuclear ribonucleoprotein H1 (H)
213738_s_at	ATP5A1	498	164360	NM_004046	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle
213757_at	EIF5A	1984	600187	NM_001970	eukaryotic translation initiation factor 5A
214617_at	PRF1	5551	170280	NM_005041	perforin 1 (pore forming protein)
214924_s_at	OIP106	22906	608112	NM_014965	OGT(O-Glc-NAc transferase)-interacting protein 106 KDa
215693_x_at	DDX27	55661		NM_017895	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27
215780_s_at	Hs 382039				
216336_x_at	AL031602				
217727_x_at	VPS35	55737	606931	NM_018206	vacuolar protein sorting 35 (yeast)
217759_at	TRIM44	54765		NM_017583	tripartite motif-containing 44
217875_s_at	TMEPAI	56937	606564	NM_020182, NM_020182, NM_199169, NM_199170	transmembrane, prostate androgen induced RNA
217917_s_at	DNCL2A	83658	607167	NM_014183, NM_014183, NM_177953	dynein, cytoplasmic, light polypeptide 2A
217933_s_at	LAP3	51056	170250	NM_015907	leucine aminopeptidase 3
218094_s_at	C20orf35	55861		NM_018478, NM_018478	chromosome 20 open reading frame 35
218237_s_at	SLC38A1	81539		NM_030674	solute carrier family 38, member 1
218242_s_at	CGI-85	51111		NM_016028, NM_016028	CGI-85 protein
218325_s_at	DATF1	11083	604140	NM_022105, NM_022105, NM_080796	death associated transcription factor 1
218345_at	HCA112	55365		NM_018487	hepatocellular carcinoma-associated antigen 112
218346_s_at	SESN1	27244	606103	NM_014454	sestrin 1
218704_at	FLJ20315	54894		NM_017763	hypothetical protein FLJ20315
218802_at	FLJ20647	55013		NM_017918	hypothetical protein FLJ20647
218898_at	CT120	79850		NM_024792	membrane protein expressed in epithelial-like lung adenocarcinoma
218943_s_at	RIG-I	23586		NM_014314	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide
218963_s_at	KRT23	25984	606194	NM_015515, NM_015515	keratin 23 (histone deacetylase inducible)
218956_at	GALNT6	11226	605148	NM_007210	UDP-N-acetyl-alpha-D-galactosamine polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6)
220658_s_at	ARNTL2	56938		NM_020183	aryl hydrocarbon receptor nuclear translocator-like 2
220951_s_at	ACF	29974		NM_014576, NM_014576, NM_138932	apobec-1 complementation factor
221516_s_at	FLJ20232	54471		NM_019008	hypothetical protein FLJ20232
221653_x_at	APOL2	23780	607252	NM_030882, NM_030882	apolipoprotein L, 2
221920_s_at	MSCP	51312		NM_016612, NM_016612	mitochondrial solute carrier protein
222244_s_at	FLJ20618	55000		NM_017903	hypothetical protein FLJ20618

Table 5. Genes used for the classification of MSS vs MSI tumors

Name	Symbol	Unigene	MSS	MSI
hepatocellular carcinoma-associated antigen 112	HCA112	Hs.12126	1261	653
metastasis-associated 1-like 1	MTA1L1	Hs.173043	45	91
chemokine (C-X-C motif) ligand 10	CXCL10	Hs.2248	104	274
heterogeneous nuclear ribonucleoprotein L	HNRPL	Hs.2730	194	630
hypothetical protein FLJ20618	FLJ20618	Hs.52184	776	388
splicing factor, arginine/serine-rich 6	SFRS6	Hs.6891	74	446
protein kinase C binding protein 1	PRKCBP1	Hs.75871	294	168

Table 6. Performance of the classifier

	Trainings set	Test set
	Errors in crossvalidation	Test errors
MSI	2.8% (n=25, range 0-6)	1.4% (n=10, range 0-4)
MSS	0.70% (n=30, range 0-3)	0.52% (n=29, range 0-2)
All	1.7% (n=55, range 1-7)	1.9% (n=39, range 0-5)

Table 7.

Sensitivity, Specificity, and Predictive Value of Test for MSS based on the eight gene Classifier				
Positive for MSS	True = (0.9948*29)=28,8492			False = (0.138*10)= 1.38
Negative for MSS	False = (0.0052*29)= 0.1508			True = (0.962*10)= 9.62
Sensitivity	28.9507/29	=	99.5%	
Specificity	9.62/10	=	96.2%	
Positive predictive value	28.8492/30.2292	=	95.4%	
Negative predictive value	9.62/9.7708	=	98.5%	

*Based on a prevalence for MSS of 85%

Table. 8

AFFYID	SYMBOL	LOCUSLINK	OMIM	REFSEQ	AFFYDESCRIPTION
206194_at	HOXC6	3223	142972	NM_004503	Homeo box C4
214868_at	PIWIL1	9271	605571	NM_004764.2	Piwi (Drosophila)-like 1
202520_s_at	MLH1	4292	120436	NM_000249.2	MutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2)
202517_at	CRMP1	1400	602462	NM_001313 2	Collapsin response mediator protein 1
205453_at	HOXB2	3212	142967	NM_002145.2	Homeo box B2 (HOXB2)
217791_s_at	PYCS	5832	138250	NM_002860.2	Pyrroline-5-carboxylate synthetase (glutamate gamma-semialdehyde synthetase) (PYCS)
202393_s_at	TIEG	7071	601878	NM_005655.1	TGFB inducible early growth response (TIEG)
218803_at	CHFR	55743	605209	NM_018223.1	Checkpoint with forkhead and ring finger domains (CHFR)
219877_at	FLJ13842	79698		NM_024645 1	Hypothetical protein FLJ13842 (FLJ13842)
202241_at	C8FW	10221		NM_025195.2	Phosphoprotein regulated by mitogenic pathways (C8FW)